DOCUMENT RESUME

ED 194 629 TM 800 761

AUTHOR Shayer, Michael
TITLE A New Approach to Data Analysis for the Construction of Piagetian Tests.
PUB DATE [80]
NOTE 22p.

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Children; *Cognitive Measurement; *Developmental Stages; Difficulty Level; Factor Analysis; Foreign Countries; *Test Construction; *Testing Problems; Test Items
IDENTIFIERS England; *Piagetian Stages; *Piagetian Tests

ABSTRACT

The special case of the Piagetian model is discussed in relation to test theory. Problems connected with the construction and analysis of a test based on Piaget and Inhelder's The Child's Construction of Quantities are presented, and related to a method of representing the item discrimination which is consonant with Piagetian theory. Loevinger and Lumsden's critique of test theory is utilized to achieve a representation of the test-space in two-dimensions. This representation allows an approach to testing for unidimensionality equivalent to that of factor-analysis and stays closer to the data. (Author/MH)

A NEW APPROACH TO DATA-ANALYSIS FOR THE CONSTRUCTION OF PIAGETIAN TESTS

MICHAEL SHAYER
Chelsea College, London University.

Dr. M. Shayer
Chelsea College
90 Lillie Road
London, S.W.6.
ENGLAND

THE PROBLEM OF 'PSYCHOMETRISING PIAGET'

The Piagetian literature of the Sixties abounds with confusions. Was the original Genevan research replicable? If so, were the criteria used and the mode of testing nevertheless too subjective? If replicable, was the construct underlying the different Genevan research problems unitary? If unitary, what was the status of the underlying model? If subjective, then by what process can the theory, if unitary, be objectified? Each of these questions points to the necessity of standardising the original Genevan procedures, and economy of research effort points to carrying this out in the form of group tests.

Yet, by the Seventies, the problem of 'psychometrising Piaget' (Tuddenham, 1970) had still not been solved. Although this was due in part to the confounding of the above questions, it was due also to the Piagetian data not fitting the conventional data-analysing methods. To carry out a survey of Piagetian developmental norms on a sample large enough to generalise from (Shayer, Küchemann & Wylam, 1976; Shayer & Wylam, 1978), and to carry out a study of the validity of Piaget's construct of formal operational thinking (Shayer, 1979) it was necessary to produce a solution. This paper could be sub-titled "The author's misfortunes in the wilderness of test-theory". One way out of that wilderness will be sketched, and illustrated in the construction of one of the Piagetian tests required.

Both the Piagetian model and the procedure for reporting research findings are quite different from those developed by the psychometric tradition

of norm-referenced tests. Although the former involves the reporting of a wide range of behaviours, in their categorisation these are collapsed under a small number of global descriptors, Early and Late Concrete (2A & 2B), or Early and Late Formal (3A & 3B). The latter imply a large number of behaviours, each represented by test items scored on a pass/fail basis, and compared with each other in test construction by item-analysis techniques which use the performance of large samples of subjects as tests of the items' validities and reliabilities. This typically results in an equal-interval scale with about sixty enumerable intervals. The advantage of this is to make the test estimates amenable to measurement theory: the disadvantage of the Piagetian clinical interview is that only the overall global assessments can be compared quantitatively. In many papers reporting low correlations between Piagetian tasks there is a curious innocence with regard to the reliability of the interview estimate or, indeed, to simple quantitative considerations such as whether the sample range is wide enough for the calculated correlation to be a true measure of the underlying association.

Questions of test procedure required to make a group-test situation equivalent to an interview, for the subject, are discussed in Shayer, Adey and Wylam (1980) and so will not be referred to here except to note that demonstration, feedback to the subject related to his own ideas, and flexible verbal communication between test administrator and group are essential. Taking these for granted, the essence of the problem for the test constructor is to represent all the behaviours mentioned in the original Genevan research as characteristic of one or another stage or sub-stage by test-items. In this way, as a first step, the validity of the ascription of behaviour to stage can be checked objectively by the performance of a suitable sample of children. But there is a problem. In the Genevan descriptions of behaviour sometimes the problem can be solved at a particular developmental stage. In this case a person who has higher levels of thinking at his disposal will always solve the problem. Suppose a test-item is only solved at the Early Formal (3A) level, but that subjects show consistent fallacious responses to

the item when they possess Late Concrete (2B) competence, and which differentiate such subjects from Early Concrete or lower stage subjects who show different fallacious responses. If the stage-behaviours are scored dichotomously (1,0), then either one is forced to lower the subject's score on the 2B items when he succeeds at the 3A stage by scoring it only as a 3A item, or one introduces false correlation into the test content by automatically adding a 2B score to the subject's sub-test totals if he shows the higher level 3A behaviour which solves the problem. To avoid the dilemma it is necessary to construct tests out of items each of which is scored only for success, and categorised at a particular level. By 'success' is meant 'true in reality' e.g. that Length is one of the variables affecting the period of a Pendulum (2B) or that to find that Weight is not effective the valid experimental method is to keep length and push constant, and take just two different weights (3A). To find items to test lower level competencies one looks for aspects of the problem(s) to which they are adequate. Thus for 2B items in Equilibrium in the Balance one chooses 2 : 1 or 3 : 1 ratios of weights or lengths from pivot. (Inhelder & Piaget, 1958). Thus each item will be labelled with the minimum stage required for success on it. Such a method of test-construction will simultaneously be true to the hierarchical developmental theory of Piaget, and at the same time allow of an experimental test of the validity of the theory. If the theory is not true the later stage items will not scale with the earlier stage items. Moreover the test-items are now amenable to all the usual item-analysis techniques, including correlation methods such as factor-analysis. In this way one can bring the constructivist theory of Piaget, which relates developing mental structures to the complexity of the relationships which they enable the subject to discover or impose upon the world, into contact with a test-theory and method of test-construction more usually associated with an empiricist or behaviourist approach to the increase of intelligence.

Discrimination diagrams

The argument of the previous section may be more easily appreciated in the example of the development of a particular Piagetian test (NFER, 1979). This was constructed for the purpose of estimating the Piagetian stages of children over a rather wide range - from Early Concrete to Early Formal operational. The subject matter was taken from The Child's Construction of Quantities (Piaget & Inhelder, 1974), and traced all the steps by which the child is eventually able to conceive of the density of substances as a weight/volume relationship. It was necessary to find some problems which are solved successfully by children at the early stage of development of concrete operational thinking (2A); some which are solved at an intermediate level (2A/2B); others which are rarely solved until the child possesses the whole structure of mental operations which Piaget describes as Late or mature concrete operational thinking (2B), and, finally, items which are not solved successfully (in the sense of trueness to reality) until the formal operational stage (up to 3A). The concepts involved are listed in Table 1.

Insert Table 1 here

Bearing in mind that the purpose of such tests is to estimate as precisely as possible the optimum present level of thinking which a child possesses (rather than making a random sample of his strategies) it is obviously necessary to choose problems which give a sharp signal. By choosing several non-redundant problems for each level the signals summate so as to increase the precision

of estimate. From an item-analysis point of view this means that facility is not the only characteristic of an item in which one is interested. The discrimination of an item measures the sharpness of its signal. Unfortunately it was soon found that the conventional discrimination indices do not give enough information as to the way in which the item behaves in the test context. One needs to know how well the item differentiates between a given level and those immediately below, irrespective of whether, for a given population sample, it happens to have a 50% facility, or a 10 or 90% facility. For this purpose the whole test sample, and all the test items may be used to examine the performance of each item.

First the test items are grouped according to levels, and each subject given an overall level assessment based on a 2/3 - success principle. Thus if there are three 2A items, and the subject succeeds on at least 2, he is capable at least of Early Concrete thinking. If he fails to reach the 2/3 criterion on any higher group of items, then he is assessed at the 2A level. If he succeeds on at least 2/3 of the 2B items also, he is assessed 2B, and so on. Then, for each item, the percentage of the subjects assessed overall as 2A who succeed on the item is calculated. The calculation is repeated for the 2B subjects, and for the subjects assessed overall at each of the other levels. Such a discrimination diagram for an item in Volume and Heaviness is given in Figure 2.

Figure 1 about here

Such a method allows direct inspection of the item's discrimination characteristics. Thus one item may be compared with another; with a fresh sample changes in the presentation or wording of an item may be compared, and slight changes in the scoring rules used to assess the overall level of the subject on the test may be compared with each other. The purpose of all the changes would be to increase the precision of each item, which is measured directly by the abruptness of the ogive. The discrimination level of the item can also be accurately gauged by the centre point of the ogive.

It may be remarked here that an empiricist skill-integrationist account of intellectual development can be distinguished experimentally from the Piagetian account of developing general structures by such test-analysis. The former should give gently increasing discrimination diagrams, since the particular order in which a given child would develop particular skills would depend on the accidents of their experience. As they get older or brighter the probability would merely increase that any child would have achieved a concept. On the Piagetian account an invariant sequence, dependant on the hierarchical development of mental structures, would give diagrams with sharp ogives, since if a child possessed a given structure there would be a very high probability that he would solve all tasks requiring that structure.

Scalability, unidimensionality, and the Loevinger test-theory

It is curious that the elegant, subtle and powerful critique made by Jane Loevinger (Loevinger 1947; 1948) of current test theory in the Forties, and the new methods which she described, should have featured so little in the research literature. Perhaps it is a rare example of a data-processing method developed in advance of its time, when no problems existed whose theoretical model required such analysis-techniques. Guttman scalogram analysis, though vigorously criticised, has fared better. One can cite the ultimate accolade of its presence in the SPSS package. But the reason for this is that it was conceived in response to attitude variables, whose implied

underlying model is strict scalability. It has been widely used in sociology rather than psychology. A bipolar variable, such as xenophobia (and xenophilia) should be expected to scale, since feelings and attitudes are usually unified. But to place a subject somewhere on a xenophobia-xenophilia scale implies a different model from that underlying any account of developing intelligence. Even the Piagetian account, which seems to come closest to implying scalability, differs importantly from attitude variables in that no successes of an earlier stage are lost with the development of a later stage. For example, simple cause and effect thinking such as the connection between the weights on a spring and its extension would still be used by a person capable of Late Formal thinking, because such Late Concrete thinking is perfectly adequate to the relationship in question. Scalogram analysis would seem best to fit data in which there are gradual qualitative changes in behaviours or strengths of feeling over the whole scale.

Loevinger's approach ran parallel to the development of factor-analytic methods. It was, in part, an attempt to develop a method of test-analysis which would ensure that the test actually measured something. Rather than produce composite intelligence tests, and find out afterwards by factor-analysis which set of abilities are estimated by the different items, she announced that it was better to start with a theory which should impose a unified construct on a test, select test items in accordance with the theory, and then use her own appropriate method of test-analysis to improve the tests. Yet to the author's knowledge this approach was never actually used. If there is a well-defined mental construct, then it should be possible to measure increasing development of it, by subjects. A unidimensional test derived from the construct should, of course, be uni-factor (Lumsden, 1961). But as will be discussed later, there are technical reasons why factor-analysis may not give a clear decision where test-item data cover a very wide range of mental functioning. Loevinger's definition of test homogeneity

allows the functioning of each item to be inspected directly. Each item should be related to another item which tests a greater degree of achievement of the underlying construct by the relationship 'if the latter, then the former' She developed three indices which quantified this relationship for one item in relation to another (H_{ij}), for an item in relation to the test as a whole, (H_{it}), and for the degree of homogeneity of the test (H_t). Again, it will be seen later that her various H indices encounter a technical problem related to the 'difficulty-factor' problem in factor-analysis. But her principle of test-analysis is unaffected by this, and is obviously a close fit to the Piagetian model. Both Nassefat (1963) and Goldschmid and Bentler (1968) have used it on Piagetian data. The discrimination-level diagrams referred to earlier are obviously closely related to the Loevinger analysis. A sharp ogive will mean a high homogeneity within the context of the test as a whole, and a set of items of different facilities, each with sharp discrimination, will define experimentally a unidimensional construct. Thus we have a method of test-analysis which should suit Piagetian data if the Piagetian model is valid. Figure 2 gives the complete set of discrimination diagrams for the test, Volume and Heaviness, determined on a representative sample of 12 year olds.

Fig.2 about here

PROBLEMS WITH FACTOR-ANALYSIS

Attempts to test for unidimensionality by factor-analysis run foul of the 'difficulty-factor' problem (Ferguson, 1941) on data such as this. One can factor-analyse the items within a composite Piagetian test, and find oneself with, not one factor as the Piagetian model requires, but a 'Concrete factor' and a 'Formal factor' and even possibly an intermediate

'Transitional formal factor' as well (Lawson 1978). Factor-analysis as a black-box tool is really best used on data all of which are around the 50 per cent facility level for the sample chosen. It will then accurately differentiate the correlation matrix into the number of factors required to explain the data. The reason for this has most clearly been shown by Carroll (1961). Factor-analysis is a process of grouping of the cells of a correlation matrix. If the correlation coefficient used is Pearson r, or the phi-coefficient which is the form it takes for dichotomous data, then the maximum value it can take is limited by the degree of overlap in the two-dimensional matrix of the variables. Ferguson showed that if a set of items span a facility range from, say, 10 to 90 per cent, the correlation matrix will split into at least three sets, even though the 'true' correlation between all the items is the same. Items with facilities in the same range may attain a value of nearly 1, if perfectly correlated, but when correlated with items in a different facility range may be limited to a maximum of 0.5 or less, and will be lowered proportionally if less than perfectly correlated. Thus the factor-analysis procedure can produce several factors from a uni-dimensional set of items.

A partial solution to this problem was offered by Bentler (1971) under the name of Monotonicity Analysis, and has been used both by him (Goldschmid & Bentler, 1968) and Hooper and Dihoff (1975) in the analysis of Piagetian data. In effect the method involves changing the association index to one which does not drop when the facility of items varies. The index he used is one proposed by Yule in 1912, which was developed by Yule in response to an analogous problem where use either of Pearson r or of tetrachoric r produced either negative or positive distortion of the association relationship when the marginal values differed widely. Yule's Y or omega index (Bentler's 'm' reduces to Y, for dichotomous data) was conceived to yield, as nearly as

TABLE 2

VOLUME AND HEAVINESS: PRINCIPAL COMPONENTS

Item	Component loadings after Varimax rotation	
	Component 1	Component 2
1		74
2		69
3a	57	36
3b		43
5	68	
6	54	
7	61	
8	67	30
9	66	
10	67	
11	34	
12	50	
13a	57	
13b	67	
14	38	

possible, the same value which the phi-coefficient would have given for the data-matrix if it had been cut to give 50/50 marginal values. Even this coefficient distorts when the value of one of the four correlation cells drops nearly to 0, but it does so less than any coefficient, including Yules Q. Yule showed that the sampling variation for this coefficient is less than that of any of the previously named variables. The reason for this can be seen in Carrol's diagrams, as can be also seen the technical reason why Loevinger's H (this can be shown to be identical to phi/phi_{max}) cannot be used as a solution to the 'difficulty-factor' problem. It, like tetrachoric r, distorts strongly, but positively, when the correlation matrix is cut at extreme values. It therefore also leads to 'difficulty-factors' - but in this case by associating items of widely <u>differing</u> facilities.

Technically, the procedure is to use a simple principal components analysis programme on a matrix of Yule's Y coefficients. It is easy to write a programme to compute Yule's Y, and insert it in the SPSS PA 1 programme in place of the phi-matrix. When this was carried out for the Volume and Heaviness task, it produced a two-component solution, as given in Table 2.

Insert Table 2 here

However, it has to be admitted that all this is stretching the factor-analysis procedure to the analysis of hierarchical data to which it is not really suited.

The problem is that the Component 2 loadings are on high facility items. Even Yule's Y cannot pick up a relationship when there is virtually no overlap in the data. As will be seen later items 1 to 3 can be accommodated quite well within the context of the overall test construct. What is required is a data-analysing technique which represents all the information in the test-data.

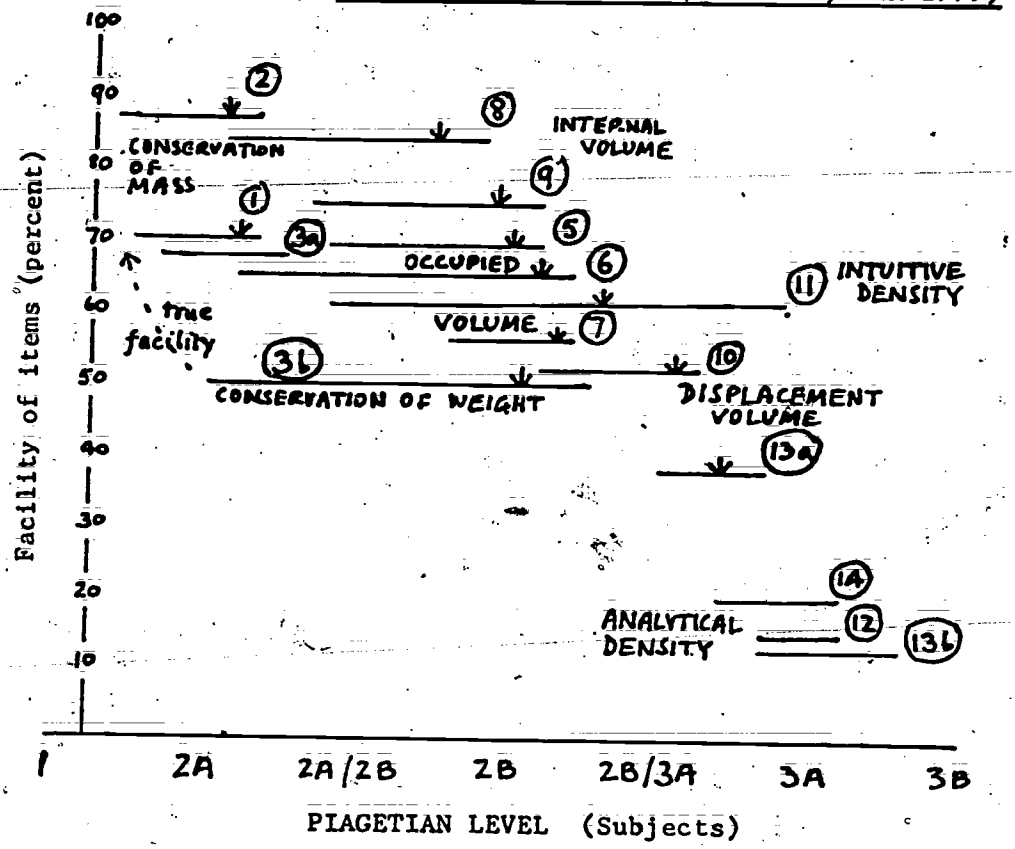
REPRESENTATION OF TEST-CONTENT

Lumsden's 'flogging-wall'

In a powerfully argued review (Lumsden, 1976) Lumsden playfully offered a concrete metaphor to try to elucidate some of the paradoxes and fallacies involved in eight years and more of the psychometric test literature. The same kind of conclusion was reached much earlier in a penetrating review called 'The Attenuation Paradox' by Jane Loevinger. She showed that there was only one kind of item distribution (Loevinger 1954, Fig.3) in which the validity did not decrease with increasing reliability:- that in which the items had rectangular distribution of item facilities. Lumsden's 'flogging-wall' suggests an explanation for this. He suggests an analogy between the test-estimation process, and an attempt to estimate the height of a subject by drawing him past a wall out of which are waving canes each at a different height, but some flogging up and down with different amplitudes. The subject's height is estimated by the number of cane-cuts he accumulates in the course of his trial. By analogy, the discrimination and spacing of the items in a test should be such as to compromise between leaving no gaps, and ensuring that at each level of the test there are enough overlapping items to increase the precision of estimate. The discrimination should not be too coarse or the reliability, or precision of estimate, will be low.

This metaphor suggested a method of representing the three important aspects of items in two dimensional space. As Loevinger (1954,p503) pointed out, one needs a quantitative estimate of the parameters of facility, discrimination level, and discrimination power, or fineness. From the discrimination diagrams used earlier one may, by imposing an equal-interval

Fig.3: Discrimination range of items in Volume and Heaviness (Task II, NFER 1979)



241 9-12 year-olds. Mixed. Slightly above average sample.

scale onto the Piagetian levels used, estimate the Piagetian levels of the subjects who, successively, show 25%, 67%, and 75% success on an item. Note that these are not facility levels. The 67% level is that at which 67 percent of the subjects assessed by the test overall at that level pass the item. In effect, this is the discrimination level* of the item. In Figure 3 the facility of the items is plotted against a line spanning the 25, 67 & 75% levels for each item. The length of the item line is an inverse measure of the unidimensionality of the item in the test context and estimates discrimination power. It will be seen that all items discriminate with a satisfactory sharpness, with the exception of item 3b, Conservation of Weight, and item 11, Intuitive Density. This, it is true, was reflected in the low communalities of these two items in the factor-analysis, but the significance is clearer in this diagram. In the case of item 11 one must say that it is not as closely related to the overall developmental construct as are the other items. Item 3b looks as though it should discriminate at a lower level, but there is clearly some aspect of the weight conservation problem (this is of a grain of corn being 'popped' by heat) which renders the facility less and the discrimination level higher than one might have expected. This points to some deficiency not picked up earlier in the formulation of the item itself. Further Piagetian research questions are obvious, but it is not the purpose of this paper to explore them.

Such a diagram does most of the work in estimating the unidimensionality of a test, and has the advantage both of representing the discrimination levels of the items and their spacing within the test, and also of pin-pointing the departures of any items from the overall test-construct. It will not work in reverse, of course. A multi-dimensional test would have item-lines all stretching widely across the test-space. Only factor-analysis would indicate how many factors were involved. But where, as in Piagetian studies, it is a unidimensional construct one is attempting to explore, such a diagram does represent all the parameters which Loevinger

was attempting to characterise in test-construction, and provides an overall check on the test construct which can suggest immediate remedy.

---

* A 67% success criterion was originally taken for pedagogic reasons. It seemed a reasonably stringent proportion by which to tell whether a pupil understood the basic principle underlying several items testing the same science concept. Subsequently it was found empirically to be the cutting level which gave the best scaling of groups of items which differed widely in facilities, and which were expected to scale at several different levels. There may be a good technical reason for this, but it is not known to the author.

Bentler, P.M. Monotonicity analysis: an alternative to linear factor and test analysis. 220-244 in Measurement and Piaget, Green D.R. (Ed) New York: McGraw-Hill, 1971.

Carroll,J.B. The nature of the data, or how to choose a correlation coefficient. Psychometrika, 26, (4), 347-372, 1961.

Ferguson, G.A. The factorial interpretation of test difficulty. Psychometrika, 6, 323-329, 1941.

Goldschmid, M.L. & Bentler, P.M. The dimensions and measurement of conservation. Child Development, 39, 787-802, 1968.

Hooper, F. & Dihoff, R.E. Multidimensional scaling of Piagetian task performance. Madison: Wisconsin Research and Development Centre for Cognitive Learning, 1975.

Inhelder, B. & Piaget, J. The Growth of Logical Thinking. London: Routledge & Kegan Paul, 1958.

Lawson, A.E. The development and validation of a classroom test of formal reasoning. Journal of Research and Science Teaching, 15, 1, 11-24, 1978.

Loevinger, J. A systematic approach to the construction and evaluation of tests of ability. Psychological Monographs, 61, No.4, 1947.

Loevinger, J. The technic of homogeneous tests compared with some aspects of 'scale analysis' and factor analysis. Psychological Bulletin, 45, 507-529, 1948.

Loevinger, J. The attenuation paradox in test theory. Psychological Bulletin, 51, 5, 493-504, 1954.

Lumsden, J. The construction of unidimensional tests. Psychological Bulletin, 58, 2, 122-131, 1961.

Lumsden, J. Test Theory. Annual Review of Psychology, Vol.27, 251-280, 1976.

Nassefat, M. Étude quantitative sur l'évolution des opérations intellectuelles. Neuchâtel: Delachaux et Niestlé, 1963.

NFER. Science Reasoning Tasks. N.F.E.R. Publishing Co., 2 Oxford Road East, Windsor, Berks., England, 1979.

Piaget, J. & Inhelder, B. The child's construction of quantities. London: Routledge & Kegan Paul, 1974.

Shayer, M. Has Piaget's construct of formal operational thinking any utility? British Journal of Educational Psychology, 49, 265-276, 1979.

Shayer, M., Adey, P. & Wylam, H. Group Tests of Cognitive Development: Ideals and a Realisation. Journal of Research and Science Teaching, 1980.

Shayer, M., Küchemann, D.E. & Wylam, H. The distribution of Piagetian stages of thinking in British middle and secondary school children. British Journal of Educational Psychology, 46, 164-173, 1976.

Shayer, M. & Wylam, H. The distribution of Piagetian stages of thinking in British middle and secondary school children. II: 14-16 year-olds and sex differentials. British Journal of Educational Psychology, 48, 62-70, 1978.

Tuddenham, R. in Dockerell, W.B. On intelligence. The Toronto Symposium on Intelligence, 1969. London: Methuen, 1970.

Yule,G. Udny. On the methods of measuring association between two attributes. Journal of the Royal Statistical Society, 75, 579-642, 1912.

TABLE 1

CONCEPTS IN VOLUME AND HEAVINESS TEST

Concepts	Piagetian level of concept	
Conservation of substance (Mass)	2A	Early Concrete
Internal volume and intuitive density (Heaviness)	2A/2B	
Conservation of weight and occupied volume	2B	Late Concrete
Displacement volume	2B/3A	
Density as a weight to volume relationship	3A	Early Formal

TABLE 2

VOLUME AND HEAVINESS: PRINCIPAL COMPONENTS

Item	Component loadings after Varimax rotation	
	Component 1	Component 2
1		74
2		69
3a	57	36
3b		43
5	68	
6	54	
7	61	
8	67	30
9	66	
10	67	
11	34	
12	50	
13a	57	
13b	67	
14	38	

Fig.1 Volume & Heaviness. Item 9.

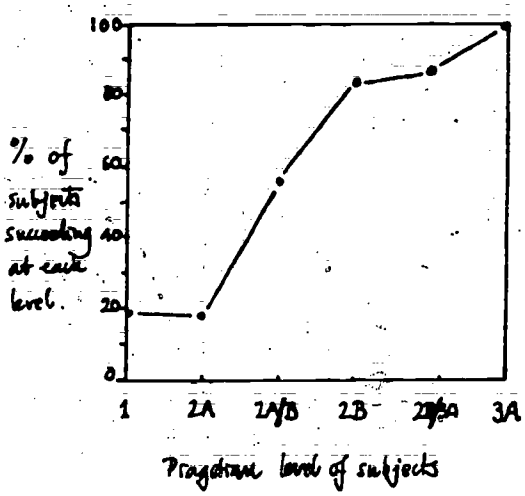


Fig.2: Discrimination-level diagrams for questions in Task II: Volume and Heaviness

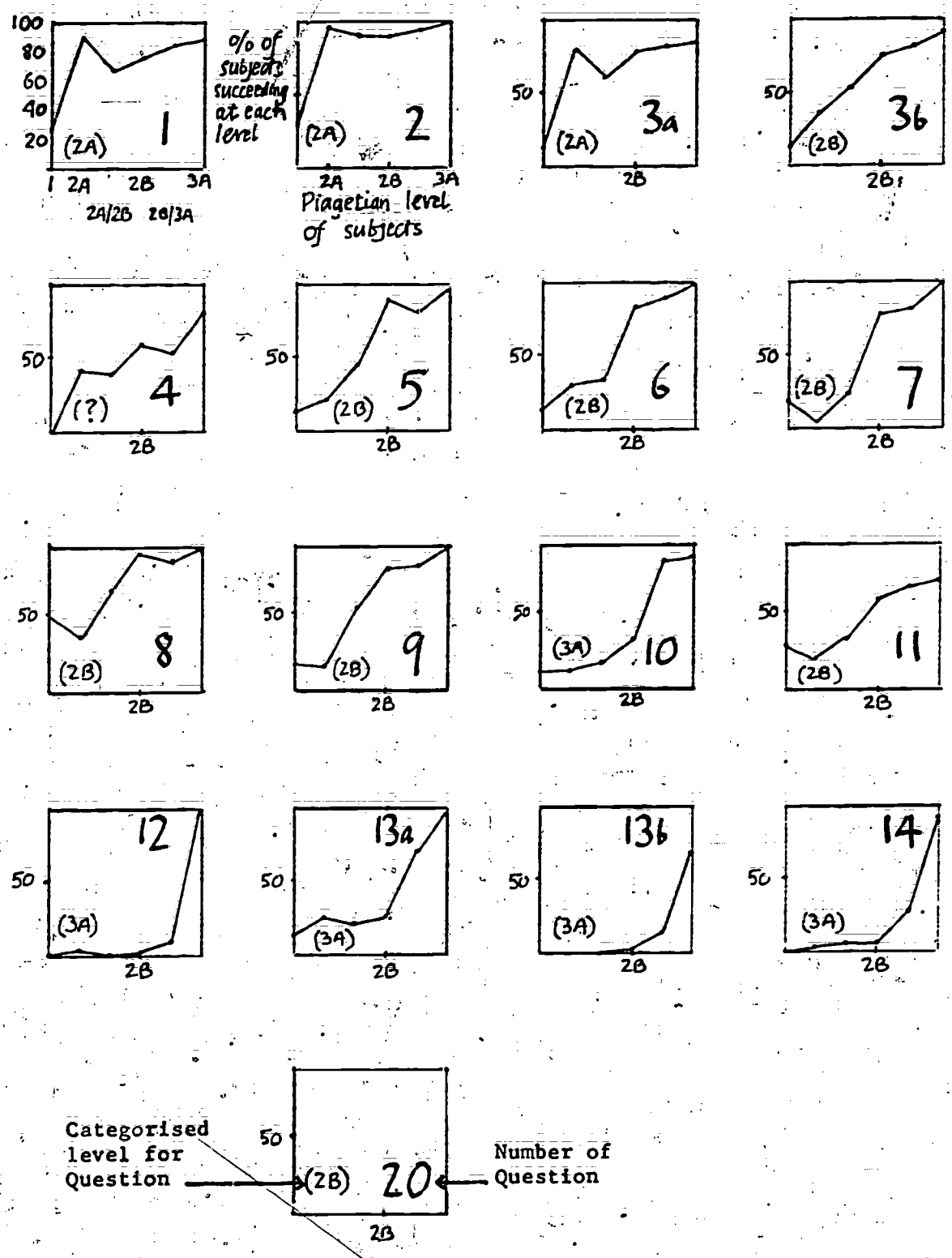
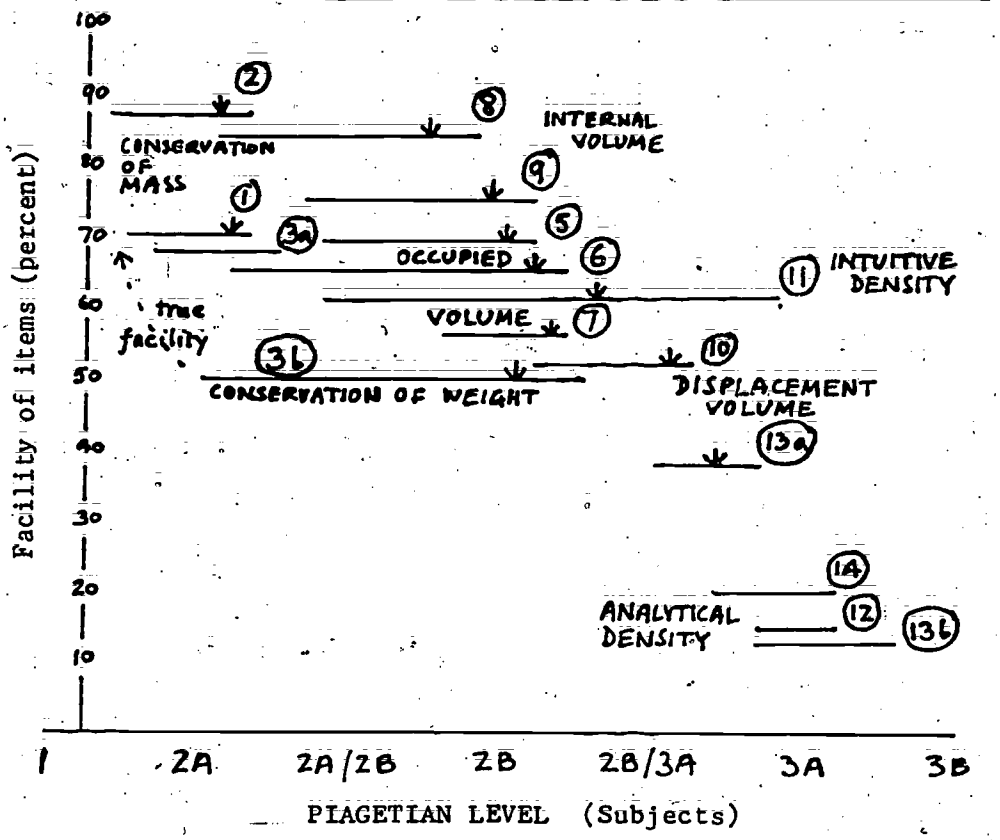


Fig. 3: Discrimination range of items in Volume and Heaviness (Task II, NFER 1979)



241 9-12 year-olds. Mixed. Slightly above average sample.